Quantifying the transcriptome of a human pathogen

Aprianto, Rieza

3

High-resolution analysis of the pneumococcal transcriptome under a wide range of infection-relevant conditions

Rieza Aprianto[a,#], Jelle Slager[a,#], Siger Holsappel[a] and Jan-Willem Veening[a,b]

[a] Molecular Genetics Group, Groningen Biomolecular Sciences and Biotechnology Institute, Centre for Synthetic Biology, University of Groningen, Nijenborgh 7, 9747 AG Groningen, the Netherlands

[b] Department of Fundamental Microbiology, Faculty of Biology and Medicine, University of Lausanne, Biophore Building, CH-1015 Lausanne, Switzerland

[#] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors

bioRxiv | https://doi.org/10.1101/283739 | 22 March 2018

Under revision for Nucleic Acids Research

RA designed the research, performed the research, analyzed the data and wrote the article.

Abstract

Streptococcus pneumoniae is an opportunistic human pathogen that typically colonizes the nasopharyngeal passage and causes lethal disease in other host niches such as the lung or the meninges. How pneumococcal genes are expressed and regulated at the different stages of its life cycle, as commensal or as pathogen, has not been entirely described. To chart the transcriptional responses of *S. pneumoniae*, we quantified the transcriptome under 22 different infection-relevant conditions. The transcriptomic compendium exposed a high level of dynamic expression and, strikingly, all annotated pneumococcal genomic features were expressed in at least one of the studied conditions. By computing the correlation of gene expression of every two genes across all studied conditions, we created a co-expression matrix that provides valuable information on both operon structure and regulatory processes. The co-expression data is highly consistent with well-characterized operons and regulons, such as the PyrR, ComE and ComX regulons, and has allowed us to identify a new member of the competence regulon. Finally, we created an interactive data center named PneumoExpress (www.veeninglab.com/pneumoexpress) that enables users to access the expression data as well as the co-expression matrix in an intuitive and efficient manner, providing a valuable resource to the pneumococcal research community.

Introduction

S. pneumoniae (the pneumococcus) is a successful opportunistic human pathogen with high carriage rates in children, immunocompromised individuals and the elderly. The pneumococcus claims the lion's share of all mortality related to LRTIs (Lower Respiratory Tract Infections), single-handedly placing LRTIs as the deadliest communicable disease[1]. Additionally, LRTIs are the second principal cause for loss of healthy life[2] (disability-adjusted life years, a combination of mortality and morbidity). Furthermore, the pneumococcus is part of the typical microbiota of the respiratory tract[3–5], with four in five young children[6] (< 5 years) and one in three adults[7] carrying the bacterium. Moreover, young children[8] and the elderly[9] are especially susceptible to pneumococcal pneumonia. In addition to lung infection, *S. pneumoniae* is responsible for other lethal infections, such as sepsis and meningitis[10,11].

Sequenced pneumococcal genomes from clinical and model strains reveal the presence of a pan-genome of approximately 3,473 genes[12], with up to 90% gene conservation between strains. On the other hand, individual pneumococcal strains have a relatively small genome with around 2 million bps. For example, strain D39, one of the work horses of pneumococcal research, has 2,046,572 bps with 2,150 updated genomic features (**Chapter 2**, https://veeninglab.com/pneumobrowse). Specifically, prokaryotic genome size strongly correlates with the number of coding sequences[13]. In effect, this limits the quantity of dedicated effectors for specific conditions. One of the strategies to circumvent this limit is for the genome to encode moonlighting proteins; for example, α-enolase, a major glycolytic enzyme, also binds human plasminogen, thereby combining carbon metabolism and cellular adhesion in one molecule[14,15]. In addition, the pneumococcus employs a strategy to optimize regulation of gene expression in response to environmental stimuli, in short, fine-tuning the timing and amount of gene products to ensure optimal survival. For example, we and others have observed the pneumococcus aptly imports varied carbon sources[16,17] and epithelium-adherent pneumococci express different carbohydrate importers than planktonic bacteria[18].

Here, we precisely quantified the transcriptome of *S. pneumoniae* (strain D39V, CP027540, **Chapter 2**) when exposed to 22 infection-relevant conditions. Next, we classified the annotated features into genes that are steadily highly expressed and genes with condition-dependent, dynamic expression. Furthermore, we generated a co-expression matrix containing the transcription correlation value of every possible pair of genes. We exploited the matrix to identify a new member of the competence regulon: a small hypothetical protein, encoded by SPV_0391 (*briC*). Furthermore, we provide the compendium consisting of normalized expression values, exhaustive fold changes and the co-expression matrix in PneumoExpress (www.veeninglab.com/pneumoexpress), a user-friendly browseable data center, enabling easy access. Finally, users can simply browse genomic environment around gene(s) of interest (via crosslink to www.veeninglab.com/pneumobrowse). The work and data presented here provide a valuable resource to the pneumococcal and microbial research community.

Results

Infection-relevant conditions: creating the compendium

The natural niche of *S. pneumoniae* is the human nasopharynx. However, due to its lifestyle, the bacteria may encounter greatly varied microenvironments. Interhost transmission, for example, involves the switch from cellular adherence in the nasopharynx to airborne respiratory or surface-associated droplets. In both cases, the bacteria must survive a lower temperature, desiccation and oxygenated air[19]. Inside the host, sites of colonization and infections are equally challenging with varying acidity and differing levels of oxygen and carbon dioxide, diverse temperatures and scarcity of carbon sources[20], not to mention the action of the innate and adaptive immune system. Additionally, the pneumococcus resides in a biodiverse niche with numerous occupants, including other pneumococcal strains, bacteria, fungi and viruses[20]. Theoretical models indicate that to colonize successfully in dynamic and competitive environments, bacteria must adapt their phenotype according to environmental demands[21]. To achieve this plasticity, flexible gene regulation and continuous

fine-tuning of gene expression are indispensable, although the extent to which the pneumococcus tunes its gene expression is unknown.

To reveal the degree of global gene regulation occurring in the pneumococcus under infection-relevant conditions, we exposed strain D39V to 22 conditions that mimic, to a certain extent, the varying host environment and quantified the resulting genome-wide transcriptional responses. The conditions and growth media were chosen to recapitulate the most relevant microenvironments *S. pneumoniae* might encounter during its opportunistic-pathogenic lifestyle (**Fig. 1A** and **Supplementary Table S1**). The main host-like growth conditions selected were: (i) nose-like (mimicking colonization, NEP), (ii) lung-like (mimicking pneumonia, LEP), (iii) blood-like (mimicking sepsis, BEP), (iv) cerebrospinal fluid-like (CSF-like; mimicking meningitis, FEP), (v) transmission-like, (vi) laboratory conditions (C+Y, CEP) that allow rapid growth and (vii) co-incubation with human lung epithelial cells. The growth medium for the first five conditions was based on a chemically defined medium[22]. In C+Y, a commonly used semi-chemically defined medium, we included three competence timepoints, 3, 10 and 20 minutes (C03, C10 and C20) after the addition of exogenous competence stimulating peptide-1 (CSP-1). Pneumococcal natural competence may be a response to living in a diverse ecosystem to combat stress and/or acquire beneficial genetic material from related strains and other bacteria[23,24]. Indeed, competence is induced by several stress factors including DNA damage[25–27] and moderately by co-incubation with epithelial cells[18]. Importantly, the competence regulon is well-characterized[28,29], allowing us to benchmark the quality of our experimental data and bioinformatics pipeline.

As *S. pneumoniae* is able to migrate between niches, we also analyzed the transcriptomes of pneumococci being transferred between conditions. Specifically, nose-like to lung-like, (i, N»L), nose-like to blood-like, (ii, N»B), nose-like to CSF-like, (iii, N»F), blood-like to C+Y, (iv, B»C), C+Y to nose-like, (v, C»N), nose-like to transmission-like for 5 minutes, (vi, NT5), nose-like to transmission-like for 60 minutes, (vii, NT60), nose-like to transmission-like for 5 minutes and back to nose-like medium for 5 minutes, (viii, NTN). Moreover, a condition mimicking meningeal fever was included (F40). Transfers were performed for only 5 minutes prior to

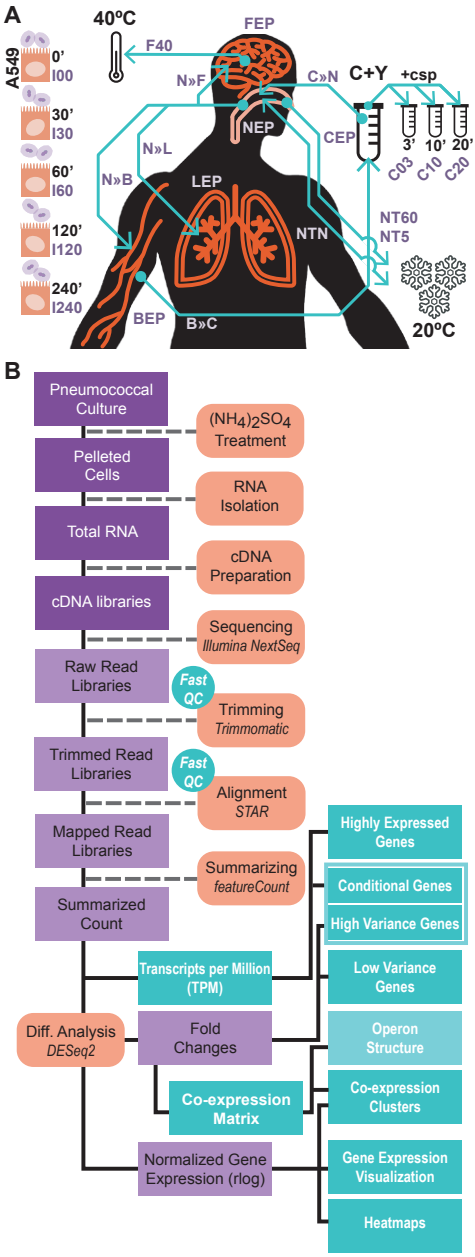


Fig. 1. Mimicking conditions relevant to the opportunistic pathogen lifestyle. **A.** 22 conditions were selected, including growth in five different media (C+Y and nose-like, lung-like, blood-like and CSF-like medium); a model of meningeal fever; eight transfers between conditions, including transmission; three competence time-points and five epithelial co-incubation time-points (**Table 2**). **B.** Total RNA was isolated after ammonium sulfate treatment to inhibit RNA degradation. cDNA libraries were prepared without rRNA depletion and sequenced. Quality control was performed before and after read-trimming. Trimmed reads were aligned and counted. Next, highly and conditionally expressed genes were categorized based on normalized read counts, while high- and low-variance genes were classified based on fold changes. High variance and conditionally expressed genes together were defined as dynamic genes.

Table 1 List of infection-relevant conditions

Conditions	Description	Libraries
NEP	Growth in nose-like medium	1,2
NT5	Growth in nose-like medium, transmission 5 min	3,4
NT60	Growth in nose-like medium, transmission 60 min	5,6
NTN	Growth in nose-like medium, transm. 5 min, nose-like medium 5 min	7,8
N»L	Growth in nose-like medium, in lung-like medium 5 min	9,10
LEP	Growth in lung-like medium	11,12
N»B	Growth in nose-like medium, in blood-like medium 5 min	13,14
BEP	Growth in blood-like medium	15,16
B»C	Growth in blood-like medium, in C+Y 5 min	17,18
N»F	Growth in nose-like medium, in CSF-like medium 5 min	19,20
FEP	Growth in CSF-like medium	21,22
F40	Growth in CSF-like medium, then 40°C (fever-like) 5 min	23,24
C»N	Growth in C+Y, in nose-like medium 5 min	25,26
CEP	Growth in C+Y	27,28
C03	Growth in C+Y, CSP 3 min	29,30
C10	Growth in C+Y, CSP 10 min	31,32
C20	Growth in C+Y, CSP 20 min	33,34
I00	Co-incubation with A549, 0 minutes post infection	35,36
I30	Co-incubation with A549, 30 minutes post infection	37,38
I60	Co-incubation with A549, 60 minutes post infection	39,40
I120	Co-incubation with A549, 120 minutes post infection	41,42
I240	Co-incubation with A549, 240 minutes post infection	43

RNA isolation because of the rapid production and turnover of bacterial transcripts[30]. Finally, five time-points of pneumococci adhering to human lung epithelial cells, previously reported by us[18], completed the array of conditions: 0, 30, 60, 120 and 240 minutes after infection (I00, I30, I60, I120 and I240). Collectively, the 22 growth and transfer conditions are referred to as "infection-relevant conditions" (**Fig. 1** and **Table 1**).

To mimic the different environments, we manipulated seven key parameters: type and amount of carbon source, level of serum albumin, level of CO_2, temperature, acidity of the medium and presence of an epithelial monolayer (**Table 2**). We manipulated carbon source because *S. pneumoniae* can utilize at least 32 different carbon sources[16], devotes a third of all transport mechanisms to import carbohydrate[17] and generates ATP exclusively from fermentation[31]. Moreover, respiratory mucus is the sole carbon source of the niches, from 1 $g \cdot l^{-1}$ in the lung[32] to 2 $g \cdot l^{-1}$ in the nasopharyngeal passage[33]. In addition, N-acetylglucosamine (GlcNAc) is the main monosaccharide in the human mucus (32% of dry weight), followed by galactose (29%), sialic acid, fucose and N-acetylgalactose[34]. On the other hand,

Table 2 Parameters in infection-relevant growth media

	Glucose ($g.l^{-1}$)	GlcNAc ($g.l^{-1}$)	Serum albumin ($g.l^{-1}$)	CO_2 (%)	Temp. (°C)	pH	Epithelial cells (A549)
Nose-like	-	1.28[33,34]	1[33]	N.D.	30[3]	7.0[38]	-
Lung-like	-	0.64[32,34]	3[32]	5	37	7.0[38]	-
Blood-like	0.9[35]	-	67[35]	5	37	7.4[35]	-
CSF-like	0.45[39]	-	0.45[39]	5	37	7.8[40]	-
Transmission	-	-	-	N.D.	20	-	-
C+Y	1.79*	-	0.73*	N.D.	37	6.8*	-
Infection	2.0[18]	-	10[18]	5[18]	37[18]	7.4[18]	+

*this study

glucose can be found in high concentrations in blood[35]. Therefore, two sources of carbon were included: GlcNAc in nose-like and lung-like conditions and glucose in blood-like, CSF-like, C+Y and infection conditions.

In addition, temperature was maintained at 37°C for all conditions except for nose-like medium (30°C)[36] since nasal temperature ranges from 30-34°C. We set fever temperature at 40°C and transmission at 20°C (room temperature). In particular, transmission was modeled by exposing the bacteria to room temperature and ambient oxygen level on a sterile surface. Confluent epithelial cells present a biotic surface that necessitates a different pneumococcal phenotype, such as biofilm formation[37]. Furthermore, the epithelial layer actively interacts with the bacteria and fine-tunes its own transcriptome[18].

The total number of trimmed reads per library ranges from 26 to 149 million reads (average: 89 million reads). Most reads from the non-depleted libraries aligned to the four rRNA loci of the pneumococcus. On average, 95.4% of reads mapped to rRNAs, ranging from 93.4 to 97.7% (**Fig. 2A**). At the same time, reads mapping to tRNAs occupied only 0.03% of total reads (0.01-0.05%) in the non-depleted libraries. Excluding rRNA and taking into account the read length (75 nt), the sequencing depth of libraries (i.e. coverage of the genome) ranges from 76 to 1944, suitably deep to elucidate gene expression[41].

Principal component analysis (PCA) indicated that three clusters can be observed that roughly correspond to the basal medium used to simulate the conditions. The first cluster consists of the five time-points during co-incubation with human epithelial cells, while the second

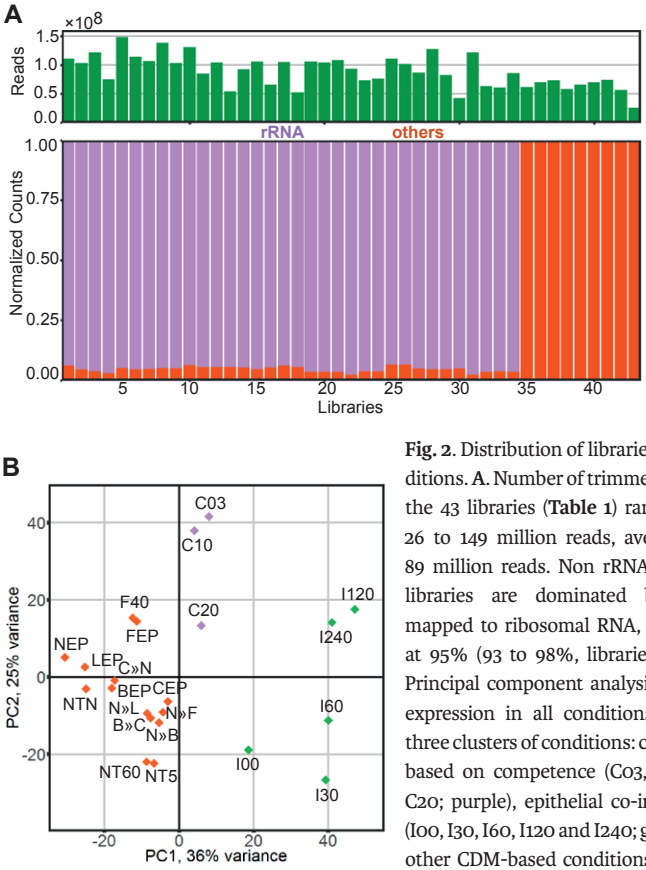


Fig. 2. Distribution of libraries and conditions. **A**. Number of trimmed reads of the 43 libraries (**Table 1**) ranges from 26 to 149 million reads, averaging at 89 million reads. Non rRNA-depleted libraries are dominated by reads mapped to ribosomal RNA, averaging at 95% (93 to 98%, libraries 1-34). **B**. Principal component analysis of gene expression in all conditions showed three clusters of conditions: conditions based on competence (C03, C10 and C20; purple), epithelial co-incubation (I00, I30, I60, I120 and I240; green) and other CDM-based conditions (orange).

cluster is made up of the competence time-points. Finally, the third cluster contains all other conditions (**Fig. 2B**). Interestingly, growth in C+Y (CEP) clusters with the latter group and not with competence samples, indicating that clustering represents biological responses, and not solely on type of medium. On the other hand, since the first cluster contained data from a different preparation and sequencing batch, we performed batch effect correction[42]. However, we failed to see an appreciable difference in distribution before and after removal of the putative batch effect and concluded that the clustering behavior is due to different biological

responses. Subsequently, downstream analysis used the non-corrected dataset. To visualize gene expression, we generated the "shortest tour" through the PCA plot (**Supplementary Fig. S1**). We calculated distances between infection-relevant conditions in the PCA plot, in an Euclidean manner and subsequently selected the minimum total distance between all the conditions using a TSP algorithm[43]. We have further validated gene expression values by qPCR (**Supplementary Fig. S2**). Taken together, we observed large differential gene expression of *S. pneumoniae* depending on its environment.

Categorization of genes: highly expressed and dynamic genes

Normalization of read counts was performed in two ways: TPM[44] (transcripts per million) and regularized logarithm[45] (rlog). While TPM-normalization corrects for the size of the library and length of a feature, rlog scales abundance directly to a $\log_2$-scale while adjusting for library size. The rlog is more suitable for visualization of gene expression across diverse conditions, while TPM values were used to categorize genes as highly or lowly expressed.

The 73 highly expressed genes, 46 of which are described as essential[46], include genes encoding rRNAs and 34 genes coding for ribosomal structural proteins. Other genes, including the two translation elongation factors *fusA* and *tuf*, DNA-dependent RNA polymerase *rpoA*, transcription termination protein *nusB*, and DNA binding protein *hlpA*, were also highly expressed in all conditions. Additionally, a set of genes associated with carbohydrate metabolism were highly expressed: *fba* (fructose-bisphosphate aldolase), *eno* (enolase), *ldh* (lactate dehydrogenase), *gap* (glyceraldehyde-3-phosphate dehydrogenase), and an ATP synthase, *atpF*. A complete list of highly expressed genes is available in **Supplementary Table S2**.

On the other hand, 48 out of the 498 conditionally expressed genes encode proteins involved in carbohydrate transport, including importers of galactosamine (*gadVWEF*), cellobiose (*celBCD*), hyaluronate-derived oligosaccharides (SPV_0293, SPV_0295-7), galactose (SPV_0559-61), ascorbic acid (*ulaABC*) and mannose (SPV_1989-92). Since we provided either N-acetylglucosamine or glucose as carbon source, the downregulation

of these sugar importers indicates that *S. pneumoniae* only expresses the importers when the target sugar is available. In contrast, some alternative sugar importers were upregulated even though in some conditions only N-acetylglucosamine or glucose was available. For example, cellobiose (*celCD*), galactose (SPV_0559-61) and multi-sugar (SPV_1583-5) transporters were upregulated upon co-incubation with epithelial cells. We postulated that the epithelial mucosal layer incites the expression of these importers since the washed epithelial layer did not result in a similar response[18]. A full list of conditionally expressed genes is available in **Supplementary Table S3**.

In addition, exhaustive comparisons (231 in total) between every set of two conditions were performed. The coefficient of variation of the summarized fold changes per gene were used to categorize high and low variance genes (**Methods**). High variance genes include pyrimidine-related genes (*pyrFE, pyrKDb, uraA, pyrRB-carAB*) and purine-associated genes (*purC, purM, purH*). These genes were activated during co-incubation (I00 to I240), transfer to transmission (NT5 and NT60) and growth in lung (LEP). Furthermore, members of the ComE regulon were heavily upregulated in all competence time points (C03, C10 and C20), CSF-like growth (FEP), fever (F40) and late co-incubation (I120 and I240). In contrast, the also dynamic expression of the ComX regulon peaks 10 minutes after addition of CSP-1 (C10) and on transfer to transmission (NT5, NT60). We have combined conditionally expressed genes and high variance genes into a single category: the dynamically expressed genes (**Fig. 3** and **Supplementary Table S4**). Visualization of low-variance genes can be observed in **Supplementary Fig. S2A**. Together, this coarse-grained analysis showed the presence of a large set of genes that are conditionally expressed (approximately 25% of all genetic features), indicating large scale rewiring of the pneumococcal transcriptome upon changing conditions.

Growth-dependent expression of rRNA

rRNA depletion introduces bias to sequenced libraries[47]. We have opted not to deplete rRNAs in the majority of the libraries, endowing the compendium with an unbiased quantification of total RNA. This approach also gave us the rare opportunity to investigate the expression levels of

ribosomal RNAs in the conditions under study. Because of rRNA abundance and stability, we adopted an alternative normalization procedure prior to calculating fold-change. Rather than normalizing rRNA read counts based on the total number of reads in the library (as is the standard procedure), we exploited read counts of low variance features to define an artificial normalization factor (**Methods**). The rRNA levels were significantly higher in fast-growing pneumococci (C+Y, CEP) compared to slow-growing cells (nose-like and lung-like growth, **Fig. 3C**). Ribosomal RNA expression in the Gram-positive model organism *Bacillus subtilis* has previously been reported to be regulated by availability of dGTP, due to the fact that the initiating nucleotide of rRNA transcripts is a GTP, rather than the more common ATP[48]. Even though rRNA operons in *S. pneumoniae* are also initiated with GTP (**Chapter 2**), we did not observe a correlation between initiating nucleotide and gene expression levels in cells growing in different media (**Supplementary Fig. S2C**). Nevertheless, in prokaryotes including *S. pneumoniae*, genes encoding ribosomal RNAs and proteins are conserved in a location close to the origin of replication [49–51]. The *ori*-proximal location of the four pneumococcal rRNA loci results in a higher gene copy number of rRNAs in fast-growing cells, such as in C+Y, as a direct consequence of the increase in replication initiation frequency. Indeed, we find that in general, constitutively expressed genes located close to the origin of replication demonstrate higher expression under fast growth[26,49].

Assembly of genome-wide correlation values to generate a co-expression matrix

We created a co-expression matrix from the fold changes between conditions. First, we exhaustively compared genome-wide fold changes between every two conditions of the 22 infection-relevant conditions. Next, we calculated the dot-product of the vector containing all fold changes of gene 1 with the vector containing all fold changes of gene 2 (*a*, non-normalized correlation value). Similarly, we determined self-dot-products of gene 1 (*b*) and gene 2 (*c*). A normalized correlation value was obtained by calculating the ratio of the non-normalized value (*a*) to the geometric mean of self-dot-products (*b* and *c*). We then mapped this correlation value according to the

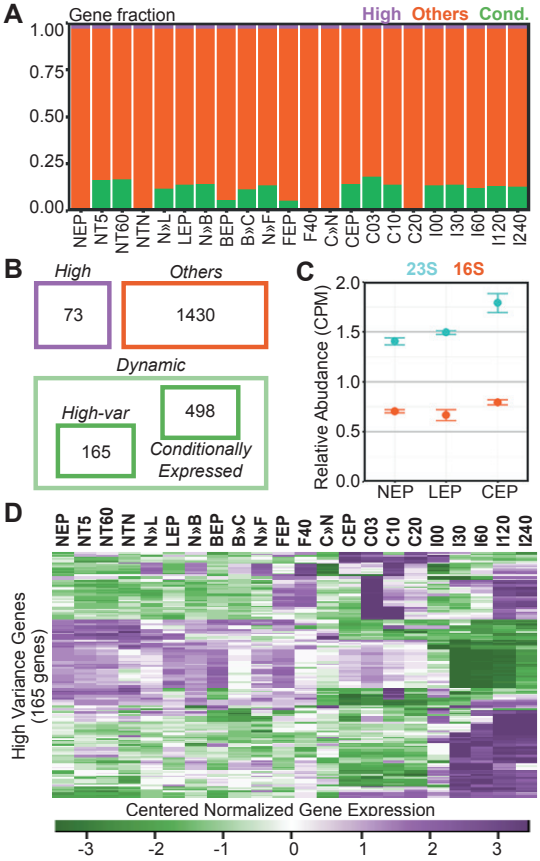


Fig. 3. Categorization of genes. **A**. Visualization of the number of genes in all conditions according to their categories: steadily highly expressed (purple), conditionally expressed (green), and others (orange). Of the 2,150 features, 73 are classified as highly expressed, while 498 features are conditionally expressed (lowly expressed in at least one condition). **B**. Highly expressed genes include essential genes, genes encoding ribosomal proteins and rRNAs. Dynamic genes are a combination of the 165 high variance genes and 498 conditionally-expressed genes. **C**. 23S rRNA was significantly downregulated in nose-like (NEP) and lung-like (LEP) growth, compared to rich C+Y growth (CEP) (p<0.05). 16S rRNA showed a similar trend but was not statistically significant (p = 0.33, CEP to NEP; p = 0.83, CEP to LEP, error bars represent standard error). **D**. Normalized expression values of high-variance genes were centered, as described in **Supplementary Methods**, and plotted as heat maps. Distinct clusters of gene expression can readily be observed (purple: high expression, green: low expression).

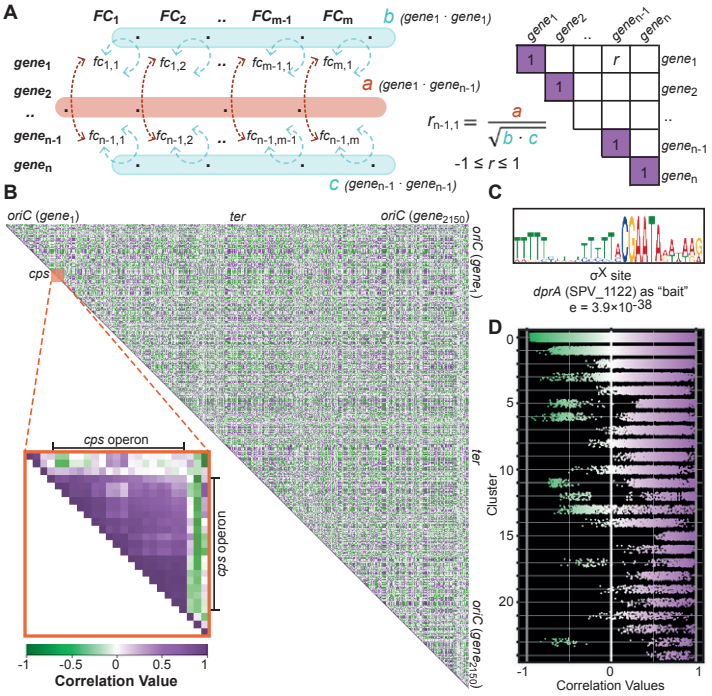


Fig. 4. Assembly of co-expression matrix from correlation values of every two pneumococcal genes. **A**. The exhaustive fold changes of every set of two genes are converted into a correlation value: first, the dot-product between two genes (*a*, orange) and the dot product of each gene with itself (*b* and *c*, blue) are calculated. The correlation value is the ratio between *a* and the geometric mean of *b* and *c*. Values were assembled by the genomic coordinates of the target genes. **B**. The co-expression matrix as a visualized gene network. Self-correlation values are 1 by definition and correlation values were plotted according to the genomic positions of target genes. Purple and green indicate positive and negative correlation values between two genes, respectively. Color intensities represent correlation strength. Blocks of highly correlated genes close to the matrix diagonal indicate operon structures, for example for the *cps* operon (inset). **C**. Enriched promoter motif recovered from genes highly correlated to *dprA* (SPV_1122) matches the consensus ComX binding site[53]. **D**. Pneumococcal genes were clustered into 25 clusters based on TPM (transcripts per million). Then, correlation values for every two genes within each cluster were plotted. Cluster 0 is non-modular and its correlation values can be considered as random. Within-cluster values showed a clear trend towards higher correlation (purple).

genomic positions of the original genes (**Fig. 4A** and **Methods**). The maximum correlation value including self-correlation is 1 and the determined correlation values range from -0.97 to 1. Close to the matrix diagonal, we observed blocks of highly correlated genes indicating their co-expression and proximity. These proximity blocks are referred to as putative operons and used as input for further analysis (**Chapter 2**). In particular, the well-known *cps* operon[52] can be observed in the co-expression matrix, whereby 16 consecutive genes are co-expressed as a single operon (**Fig. 4B** inset). In contrast, the correlation values between members of the *cps* operon and either genes upstream or downstream of the locus, are considerably lower.

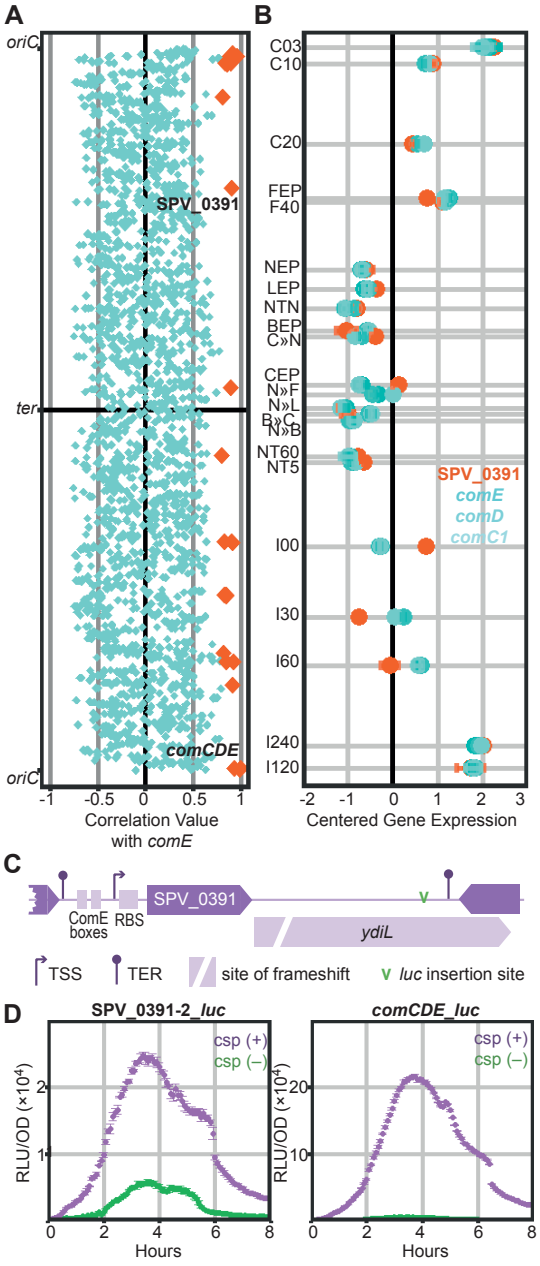
In addition to belonging to the same operon, co-expression can be mediated by shared expression-regulatory properties. Regulatory proteins typically interact with the promoter regions of regulated genes. From the matrix, we recovered 46 features (of 26 operons) that are highly correlated to *dprA,* a member of the ComX regulon. Motif enrichment analysis on the 50-nt region upstream of the corresponding 26 start sites resulted in a 28 nucleotide motif (**Fig. 4C**) that closely matches the ComX binding site as previously reported[53]. Furthermore, we clustered pneumococcal genes based on their normalized expression value (transcripts per million, TPM), and recovered 25 clusters[54,55]. The first cluster, cluster 0, is a non-modular cluster, containing all genes that did not fit into any of the other clusters. This cluster can therefore be considered as a random control. When we plotted correlation values of every set of two genes within each cluster, we observed a bias towards higher correlation values in all clusters except for the non-modular cluster (**Fig. 4D**). As an additional control, we selected 120 random genes, divided into 3 groups and plotted the correlation values within the groups. There, we observed a truly random distribution of correlation values in all groups (**Supplementary Fig. S3A**). We concluded that the co-expression matrix represents a simple network of genome-wide expression profiles that reveals meaningful transcriptomic responses to a changing environment. Moreover, by comparing gene expression profiles across a wide range of conditions, it unveils direct and indirect regulatory connections between genes. The co-expression matrix also has the potential to elucidate negative regulators by strong negative correlation coefficients with their target genes.

Exploiting the matrix to reveal a new member of the competence regulon

Two-component regulatory systems (TCSs) are essential for the pneumococcus to sense its microenvironment and to fine-tune its gene expression[56,57]. ComDE, the best-described TCS, is controlled by a quorum-sensing mechanism and regulates the activation and synchronization of competence, or X-state, which in turn is responsible for the expression of ~100 genes and a wide range of phenotypic changes[57,58]. From the co-expression matrix, we recovered genes strongly correlated with *comE*, encoding the DNA-binding regulator. Specifically, we identified 26 *comE*-associated genes with correlation values above 0.8. ComE autoregulates its own expression along with expression of *comC1* (SPV_2065) and *comD* (SPV_2064), which belong to the same operon and indeed correlate strongly with *comE*. Furthermore, other known members of the ComE regulon, such as *comAB* (SPV_0049-50), *comW* (SPV_0023) and *comM* (SPV_1744) belong to the same cluster.

Interestingly, SPV_0391, encoding a conserved hypothetical protein, was included in the group. SPV_0391 has not been reported as part of the competence regulon in array-based pneumococcal competence studies[28,29]. Furthermore, *comE*-associated genes are not localized in a specific genomic location, but spread out throughout the genome (**Fig. 5A**), ruling out the effect of genomic location. Expression values of *comCDE* and SPV_0391 across infection-relevant conditions demonstrated strong correlation between the genes (**Fig. 5B**). In the promoter region of SPV_0391,

Fig. 5. The co-expression matrix reveals a new competence-regulated gene. **A**. The gene encoding a pneumococcal response regulator, ComE, was used to recover 26 highly correlated features (orange diamonds). The group is mainly populated by known members of the ComE regulon, except for SPV_0391, a conserved hypothetical gene not previously reported to be part of the competence regulon. **B**. Centered expression values of SPV_0391 (orange) and *comCDE* (shades of blue) were plotted against the shortest tour of infection-relevant conditions. Expression values of the four genes closely clustered together. **C**. Genomic environment of SPV_0391 with two preceding ComE boxes. SPV_0391 shared operon structure to a pseudogene, *ydiL*. **D**. Firefly luciferase (*luc*) was transcriptionally fused downstream of SPV_0391 or *comCDE* to characterize their expression profiles with and without the addition of exogenous CSP-1 (competence stimulating peptide-1, 100 ng·ul^{-1}). Addition of exogenous CSP-1 incited similar luminescence profiles in SPV_0391-SPV_0392-*luc* and in *comCDE-luc* strains.


A
oriC
ter
oriC
SPV_0391
comCDE
-1
-0.5
0
0.5
1
Correlation Value
with comE
B
C03
C10
C20
FEP
F40
NEP
LEP
NTN
BEP
C»N
CEP
N»F
N»L
B»C
N»B
NT60
NT5
I00
I30
I60
I240
I120
SPV_0391
comE
comD
comC1
-2
-1
0
1
2
3
Centered Gene Expression
C
SPV_0391
ComE
boxes
RBS
ydiL
TSS
TER
site of frameshift
luc insertion site
D
SPV_0391-2_luc
comCDE_luc
csp (+)
csp (–)
RLU/OD (×10⁴)
Hours
0
2
4
6
8
10
20

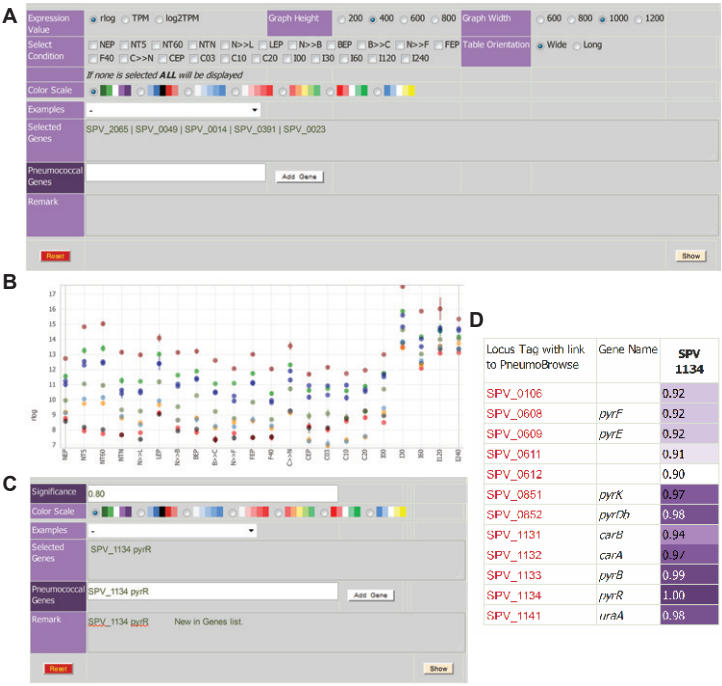


Fig. 6. An intuitive interactive database to access expression and correlation data. **A.** Users can specify their gene(s) of interest in the field "Pneumococcal Genes" and adjust other settings including normalization method, conditions to display, color scales and graph dimensions. Multiple genes of interest in a query is possible by separating names or tags by "|". Immediate genomic environment of gene(s) of interest can be explored in PneumoBrowse by clicking locus tag on the result table. **B.** Target expression values are plotted against infection-relevant conditions and the values can be downloaded for further analysis. The example shown is the nine genes correlated to *pyrR*. **C.** The co-expression matrix can be mined by simple inquiry of a gene of interest. Additionally, users can specify a desired threshold of co-expression values. **D.** Correlation values to *pyrR,* note that self-correlation is 1. Here, genomic environment can be browsed by clicking the locus tag in the result table.

we observed a ComE-binding site consisting of two ComE-boxes, which suggests direct regulation by ComE. In order to study the expression of SPV_0391 and the responsiveness of the identified ComE site, we transcriptionally inserted firefly luciferase (*luc*) downstream of SPV_0391,

which is immediately followed by pseudogene *ydiL* (SPV_2157). Importantly, no terminators or additional transcription start sites were detected between SPV_0391 and *ydiL*, suggesting they form an operon together. Previous annotation include the presence of a small hypothetical protein, SPD_0392 within *ydiL*. Thus, we chose to integrate *luc* downstream of SPV_0392 to avoid potential downstream effects (**Fig. 5C**). We compared the luminescence signal in this strain to that in a D39 derivative that expresses *luc* transcriptionally fused to the 3'-end of *comCDE*[26].

Exogenous addition of 100 ng·ul^{-1} CSP-1 stimulated luciferase activity in both strains with *luc* behind SPV_0391 and *comCDE* (**Fig. 5C**). Although the luciferase signal from SPV_0391 was an order of magnitude lower than luminescence driven by *comCDE,* the signal profiles were very similar. Difference in signal strength may stem from a weaker promoter driving SPV_0391 than *comCDE*. Additionally, we added CSP-1 after 2 hours of incubation and observed identical luminescence responses (**Supplementary Fig. S3B**). Furthermore, we disrupted SPV_0391 to elucidate its role in pneumococcal competence. Deletion of this conserved feature did not affect growth in C+Y or the expression profiles of luciferase downstream of *comCDE* and *ssbB*, member of the ComX regulon (*not shown*). Finally, transformation efficiency in the deletion strain was not significantly different from that in the parental strain. Thus, while SPV_0391 is under the control of ComE and part of the pneumococcal competence regulon, we could not determine its role in pneumococcal competence. Indeed, recent work has shown that SPV_0391 (renamed to *briC*) does not play a role in transformation but rather promotes biofilm formation and nasopharyngeal colonization (bioRxiv: doi: https://doi.org/10.1101/245902).

Development of an interactive data center to explore gene expression and correlation

To enable users to easily mine the rich data produced here, we developed an interactive data center accessible from www.veeninglab.com/pneumoexpress where users can easily extract expression values and fold changes of a gene of interest, as well as quantitative information on how its expression profile correlates with that of other genomic features (**Fig. 6**). As a proof of principle, in addition to the competence regulon,

we demonstrate results obtained by looking at the PyrR regulon. Traditional transcription factors bind the promoter region of a DNA molecule and to confidently predict all of their binding sites is challenging. PyrR, on the other hand, controls expression of its regulon by interaction with a riboswitch[59,60]. We identified four of these riboswitches (in front of *uraA*, *pyrFE*, *pyrRB-carAB* and *pyrKDb*) that are predicted to regulate the expression of nine genes, based on putative operon structures (**Chapter 2**). As expected, the eight genes show strong correlation with *pyrR* (> 0.9).

Discusssion

Extensive mineable transcriptome databases only exist for a few model bacteria such as *Bacillus subtilis*[61,62], *Staphylococcus aureus*[63], *Escherichia coli*[64,65] and *Salmonella enterica* serovar Typhimurium[66]. These resources have been proven to be invaluable for the research community. Here, we set out to map the transcriptome landscape of the important opportunistic human pathogen *S pneumoniae*. In this study, we coupled exposure to wide-ranging and dynamic infection-relevant conditions (**Table 2** and **Fig. 1A**) to high-throughput RNA-seq and generated a compendium of the pneumococcal transcriptome. Indeed, our data demonstrates that *S. pneumoniae* has a highly dynamic transcriptome with all of its genomic features differentially expressed under one conditions or the other (**Figs. 2B** and **3C**).

Previously, bacterial exposure to conditions relevant to its natural lifestyle has been reported to incite genome-wide transcriptional responses[62,67–69]. Moreover, we have shown that under these varied infection-relevant conditions, a subset of genes was constantly highly expressed while there is no gene that is always lowly expressed - highlighting the saturated and dynamic nature of the pneumococcal transcriptome. Previously, we have reported that all pneumococcal genes are expressed during early infection[18]. In this study, we again observed that throughout the array of conditions, there are no genes that are consistently silent.

The pneumococcus occupies a rich and diverse niche of the respiratory tract[20]. While we tried to estimate the relevant conditions for the

pneumococcus during its pathogenic lifestyle, other important physicochemical parameters, like the concentration of metal ions, play important roles in survival[70] and virulence[71]. Moreover, the pneumococcus shares a busy ecosystem in the respiratory tract with other bacteria, fungi and viruses[20]. Activities of other residents may be detrimental to the pneumococcal survival, as in the case of *Haemophilus influenzae* recruiting host cells to remove *S. pneumoniae*[72]. On the other hand, pneumococcal interactions with influenza viruses yield bountiful nutrients to support pneumococcal expansion[73]. Dual transcriptomics studies involving the interaction with other relevant species will offer interesting insights into pneumococcal gene expression and will greatly enhance our understanding of pneumococcal biology and pathogenesis[18,74].

Additionally, we have proposed a simple and straightforward manner to convert the dense and substantial sequencing data into a form of gene network which we call the co-expression matrix (**Fig. 4**). The matrix was assembled by arranging correlation values between two genes by their respective genomic locations. The potential of the matrix was demonstrated by the elucidation of a new member of the ComE regulon (**Fig. 5**). The gene had not been identified by array-based transcriptomics studies on the development of competence[28,29], confirming the power of next-generation sequencing over hybridization technology. Lastly, we provide the comprehensive and rich dataset to the research community by building a user-friendly online database, PneumoExpress (www.veeninglab.com/pneumoexpress) where users can easily extract expression values and fold changes of a gene of interest, as well as quantitative information on how its expression profile correlates with that of other genomic features (**Fig. 6**). By a simple click in the database, users can explore immediate genomic environments of genes of interest in PneumoBrowse. Finally, we invite other researchers to harness these resources and generate their own hypotheses, to gain new insights into pneumococcal biology and, ultimately, to identify novel treatment and prevention strategies against pneumococcal disease. In addition, the resources assist efforts in comparative genomics and transcriptomics to other bacteria.

Methods

Culturing of *S. pneumoniae* D39 and pneumococcal transformation

S. pneumoniae was routinely cultured without antibiotics. Strain construction and preparation of chemically-defined media (CDMs) are described in detail in the **Supplementary Methods**. Oligonucleotides are listed in **Supplementary Table S10** while bacterial strains are listed in **Supplementary Table S11**.

Infection-relevant growth and shock cultures of *S. pneumoniae*

The infection-relevant conditions were selected from a subset of microenvironments that the pneumococcus encounters during its opportunistic pathogenic lifestyle. We manipulated sugar type and concentration, protein level, temperature, partial CO_2 pressure and medium pH. Sicard's defined medium was selected as the backbone of infection-relevant conditions[22]. Rich C+Y medium was used for competence related conditions (**Chapter 2**) while co-incubation with epithelial cells was performed as previously described[18]. For a full description of infection-relevant conditions, *see* **Supplementary Methods**. A complete list of medium components is available in **Supplementary Table S1**.

Total RNA isolation, library preparation and sequencing

Pneumococcal cultures from infection-relevant conditions were pretreated with ammonium sulfate to terminate protein-dependent transcription and degradation. Total RNA was isolated and cDNA libraries were created without rRNA depletion. The libraries were then sequenced on Illumina NextSeq 500 as described previously[18].

Data analysis and categorization of genes

Quality control was performed before and after trimming. Trimmed reads were aligned to the recently sequenced genome (Accession: CP027540) and counted according to the corresponding annotation file, excluding 264 features that are contained by annotated pseudogenes (**Chapter 2**). Reads

were normalized into TPM[44] (transcripts per million) and by regularized log[45]. Highly expressed and lowly expressed genes were categorized from rRNA-excluded TPM. Decile values were used to partition expression values into 10 classes. The ninth decile serves as the minimum value for highly expressed genes while the first decile was used as the maximum limit for lowly expressed genes. 61 genes had TPM values above the high-limit in all infection-relevant conditions. Along with the 12 rRNA loci, these 73 genes were categorized as highly expressed genes. On the other hand, no gene is below the lower threshold in all conditions. However, 498 genes have TPM below the limit in at least one condition; the genes were categorized as conditionally-expressed.

Exhaustive fold changes were calculated[45] for every pair of two conditions out of the 22 infection-relevant conditions. Then, fold changes with low mean normalized count were set to 0. Low mean normalized count was signified by DESeq2 with "NA" as adjusted p-value. We used the formal definition of low count as previously defined[45]. Conditionally-expressed genes were excluded from the calculation of the limits of high and low variance genes because, by definition, those genes are biased towards higher variance. The coefficient of variance (cvar) for every gene across all fold changes was calculate and used as the base for variance-based partition. The cvar ninth decile was chosen as the minimum value for high variance genes, while the first decile represented the maximum limit for low variance genes. There were 165 genes with high variance, which, together with conditionally-expressed genes, were categorized as dynamic genes.

Calculations of rRNA fold changes required an alternative approach since normalization based on library size cannot be used on highly abundant features such as rRNAs. Instead, the expression values of the least variable half of all genes (1,071 features) was used as normalization factor for rRNA expression values[75]. Then, fold changes and normalized expression values were calculated.

Generation of co-expression matrix

The genome-wide exhaustive fold changes were used to calculate the correlation value of every possible set of two annotated features. First, the dot-products between fold changes of the two target genes and

self-dot-products of each gene were calculated. Next, the dot-products were summed: between two target genes (*a*) and self-products (*b* and *c*). The summed dot-product was referred to as non-normalized correlation value. This value was normalized by calculating the ratio between the non-normalized value (*a*) and the geometric mean of summed fold-changes (*b* and *c*). In turn, the geometric mean of summed fold changes was calculated as the square root of the multiplication product between the summed self-products. The normalized correlation value was then mapped into the matrix by the genomic positions of both genes (**Fig. 4A**).

Online compendium

The compendium can be accessed at www.veeninglab.com/pneumoexpress. The data are stored in a MySQL database as gene expression values. Gene expression graphs are generated by D3 (Data Driven Documents, https://d3js.org). Gene expression is presented in DESeq2-normalized values, rlog[45], TPM[44] (transcripts per million or log-transformed TPM). Exhaustive fold changes and correlation values were included as part of the pneumococcal compendium.

Luciferin assays

Firefly luciferase (*luc*) was transcriptionally fused to the 3'-end of target operons, *comCDE* and SPV_0391-2157, to monitor gene expression levels. A kanamycin resistance cassette under a constitutive promoter was used as selection marker. Plate assays were performed in C+Y with 0.25 $mg \cdot ml^{-1}$ luciferin and with and without the addition of 100 $ng \cdot ul^{-1}$ CSP-1 from the beginning of the experiment or after 2 hours incubation.

Acknowledgements

We are grateful to V. Benes and B. Haase (GeneCore, EMBL, Heidelberg) for their continuing support in sequencing; C.J. Albers, B. Jayawardhana, E.C. de Wit and M.H. Silvis for many fruitful discussions; A. de Jong for bioinformatics support; and A. Lun (Cambridge) for insightful recommendations concerning rRNA analysis. We would like to thank the Center for Information Technology

of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster. We appreciate the following creators: Hyhyhehe, Misha Petrishchev, Alberto Gongora, Hea Poh Lin, and Icon 54 for making their cliparts freely available at thenounproject.com.

Funding

Work in the Veening lab is supported by the Swiss National Science Foundation (project grant 31003A_172861), a VIDI fellowship (864.11.012) of the Netherlands Organization for Scientific Research (NWO-ALW), a JPIAMR grant (50-52900-98-202) from the Netherlands Organisation for Health Research and Development (ZonMW) and ERC starting grant 337399-PneumoCell.

Availability of data and materials

The source code for the online compendium is available in Zenodo, https://doi.org/10.5281/zenodo.1157923. Licensed under Creative Commons Attribution-Non Commercial. The transcriptomic datasets are available in the GEO repository: accession number GSE108031.

Authors' contributions

RA and JWV designed the research, analyzed the data, and wrote the article. RA performed research. JS analyzed the data. SH built the online database. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Troeger, C. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Infect. Dis.* **17,** 1133–1161 (2017).
2. Kassebaum, N. J. *et al.* Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388,** 1603–1658 (2016).
3. Miller, E., Andrews, N. J., Waight, P. A., Slack, M. P. & George, R. C. Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: an observational cohort study. *Lancet Infect. Dis.* **11,** 760–768 (2011).
4. Bosch, A. A. T. M. *et al.* Development of upper respiratory tract microbiota in infancy is affected by mode of delivery. *EBioMedicine* **9,** 336–345 (2016).
5. Bosch, A. A. T. M. *et al.* Nasopharyngeal carriage of *Streptococcus pneumoniae* and other bacteria in the 7th year after implementation of the pneumococcal conjugate vaccine in the Netherlands. *Vaccine* **34,** 531–539 (2016).
6. Regev-Yochay, G. *et al. Streptococcus pneumoniae* carriage in the Gaza Strip. *PLoS ONE* **7,** e35061 (2012).
7. Wyllie, A. L. *et al.* Molecular surveillance on *Streptococcus pneumoniae* carriage in non-elderly adults; little evidence for pneumococcal circulation independent from the reservoir in children. *Sci. Rep.* **6,** 34888 (2016).
8. Wardlaw, T., Salama, P., Johansson, E. W. & Mason, E. Pneumonia: the leading killer of children. *Lancet Lond. Engl.* **368,** 1048–1050 (2006).
9. Welte, T., Torres, A. & Nathwani, D. Clinical and economic burden of community-acquired pneumonia among adults in Europe. *Thorax* **67,** 71–79 (2012).
10. Henriques-Normark, B. & Tuomanen, E. I. The pneumococcus: epidemiology, microbiology, and pathogenesis. *Cold Spring Harb. Perspect. Med.* **3,** (2013).
11. O Brien, K. L. *et al.* Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet Lond. Engl.* **374,** 893–902 (2009).
12. Donati, C. *et al.* Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol.* **11,** R107 (2010).
13. Hou, Y. & Lin, S. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for Dinoflagellate genomes. *PLoS ONE* **4,** (2009).
14. Bergmann, S., Rohde, M., Chhatwal, G. S. & Hammerschmidt, S. alpha-Enolase of *Streptococcus pneumoniae* is a plasmin(ogen)-binding protein displayed on the bacterial cell surface. *Mol. Microbiol.* **40,** 1273–1287 (2001).
15. Bergmann, S. *et al.* Identification of a novel plasmin(ogen)-binding motif in surface displayed alpha-enolase of *Streptococcus pneumoniae*. *Mol. Microbiol.* **49,** 411–423 (2003).
16. Bidossi, A. *et al.* A functional genomics approach to establish the complement of carbohydrate transporters in *Streptococcus pneumoniae*. *PLoS ONE* **7,** e33320 (2012).
17. Buckwalter, C. M. & King, S. J. Pneumococcal carbohydrate transport: food for thought. *Trends Microbiol.* **20,** 517–522 (2012).

18. Aprianto, R., Slager, J., Holsappel, S. & Veening, J.-W. Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. *Genome Biol.* **17,** 198 (2016).
19. Walsh, R. L. & Camilli, A. *Streptococcus pneumoniae* is desiccation tolerant and infectious upon rehydration. *mBio* **2,** e00092-11 (2011).
20. Man, W. H., de Steenhuijsen Piters, W. A. A. & Bogaert, D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat. Rev. Microbiol.* **15,** 259–270 (2017).
21. Lymbery, A. J. Niche construction: evolutionary implications for parasites and hosts. *Trends Parasitol.* **31,** 134–141 (2015).
22. Paixao, L. *et al.* Host glycan sugar-specific pathways in *Streptococcus pneumoniae*: galactose as a key sugar in colonisation and infection. *PloS One* **10,** e0121042 (2015).
23. Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol.* **6,** 288–301 (2008).
24. Veening, J.-W. & Blokesch, M. Interbacterial predation as a strategy for DNA acquisition in naturally competent bacteria. *Nat. Rev. Microbiol.* **15,** 621–629 (2017).
25. Prudhomme, M., Attaiech, L., Sanchez, G., Martin, B. & Claverys, J.-P. Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*. *Science* **313,** 89–92 (2006).
26. Slager, J., Kjos, M., Attaiech, L. & Veening, J.-W. Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell* **157,** 395–406 (2014).
27. Stevens, K. E., Chang, D., Zwack, E. E. & Sebert, M. E. Competence in *Streptococcus pneumoniae* is regulated by the rate of ribosomal decoding errors. *mBio* **2,** e00071-11 (2011).
28. Dagkessamanskaia, A. *et al.* Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae*: competence triggers stationary phase autolysis of ciaR mutant cells. *Mol. Microbiol.* **51,** 1071–1086 (2004).
29. Peterson, S. N. *et al.* Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays. *Mol. Microbiol.* **51,** 1051–1070 (2004).
30. Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. & Rosenow, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **13,** 216–223 (2003).
31. Tettelin, H. *et al.* Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293,** 498–506 (2001).
32. Kaliner, M. *et al.* Human respiratory mucus. *Am. Rev. Respir. Dis.* **134,** 612–621 (1986).
33. Ugwoke, M. I., Agu, R. U., Verbeke, N. & Kinget, R. Nasal mucoadhesive drug delivery: background, applications, trends and future perspectives. *Adv. Drug Deliv. Rev.* **57,** 1640–1665 (2005).
34. Xia, B., Royall, J. A., Damera, G., Sachdev, G. P. & Cummings, R. D. Altered O-glycosylation and sulfation of airway mucins associated with cystic fibrosis. *Glycobiology* **15,** 747–775 (2005).
35. Krebs, H. A. Chemical composition of blood plasma and serum. *Annu. Rev. Biochem.* **19,** 409–430 (1950).
36. Lindemann, J., Leiacker, R., Rettinger, G. & Keck, T. Nasal mucosal temperature during respiration. *Clin. Otolaryngol. Allied Sci.* **27,** 135–139 (2002).

37. Marks, L. R., Parameswaran, G. I. & Hakansson, A. P. Pneumococcal interactions with epithelial cells are crucial for optimal biofilm formation and colonization in vitro and in vivo. *Infect. Immun.* **80,** 2744–2760 (2012).
38. Lai, S. K., Wang, Y.-Y. & Hanes, J. Mucus-penetrating nanoparticles for drug and gene delivery to mucosal tissues. *Adv. Drug Deliv. Rev.* **61,** 158–171 (2009).
39. Deisenhammer, F. *et al.* Guidelines on routine cerebrospinal fluid analysis. Report from an EFNS task force. *Eur. J. Neurol.* **13,** 913–922 (2006).
40. Weed, L. H. The cerebrospinal fluid. *Physiol. Rev.* **2,** 171–203 (1922).
41. Creecy, J. P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Curr. Opin. Microbiol.* **23,** 133–140 (2015).
42. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
43. Dolan, E. D. NEOS Server 4.0 Administrative Guide. *arXiv* (2001).
44. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci. Theor. Den Biowissenschaften* **131,** 281–285 (2012).
45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
46. Liu, X. *et al.* High-throughput CRISPRi phenotyping identifies new essential genes in *Streptococcus pneumoniae*. *Mol. Syst. Biol.* **13,** (2017).
47. Lahens, N. F. et al. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15,** R86 (2014).
48. Krasny, L. & Gourse, R. L. An alternative strategy for bacterial ribosome synthesis: *Bacillus subtilis* rRNA transcription regulation. *EMBO J.* **23**, 4473–4483 (2004).
49. Couturier, E. & Rocha, E. P. C. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* **59**, 1506– 1518 (2006).
50. Slager, J. & Veening, J.-W. Hard-wired control of bacterial processes by chromosomal gene location. *Trends Microbiol.* **24,** 788–800 (2016).
51. Soler-Bistué, A., Timmermans, M. & Mazel, D. The proximity of ribosomal protein genes to oriC enhances *Vibrio cholerae* fitness in the absence of multifork replication. *mBio* **8**, e00097-17 (2017)."
52. Wen, Z., Liu, Y., Qu, F. & Zhang, J.-R. Allelic variation of the capsule promoter diversifies encapsulation and virulence In *Streptococcus pneumoniae*. *Sci. Rep.* **6,** (2016).
53. Luo, P. & Morrison, D. A. Transient association of an alternative sigma factor, ComX, with RNA polymerase during the period of competence for genetic transformation in *Streptococcus pneumoniae*. *J. Bacteriol.* **185,** 349–358 (2003).
54. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
55. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J. Stat. Softw.* **46,** (2012).
56. Gomez-Mejia, A., Gamez, G. & Hammerschmidt, S. *Streptococcus pneumoniae* two-component regulatory systems: The interplay of the pneumococcus with its environment. *Int. J. Med. Microbiol.* (2017). doi:10.1016/j.ijmm.2017.11.012

57. Havarstein, L. S., Hakenbeck, R. & Gaustad, P. Natural competence in the genus Streptococcus: evidence that streptococci can change pherotype by interspecies recombinational exchanges. *J. Bacteriol.* **179,** 6589–6594 (1997).
58. Moreno-Gamez, S. *et al.* Quorum sensing integrates environmental cues, cell density and cell history to control bacterial competence. *Nat. Commun.* **8,** 854 (2017).
59. Bonner, E. R., D'Elia, J. N., Billips, B. K. & Switzer, R. L. Molecular recognition of pyr mRNA by the *Bacillus subtilis* attenuation regulatory protein PyrR. *Nucleic Acids Res.* **29,** 4851–4865 (2001).
60. Martinussen, J., Schallert, J., Andersen, B. & Hammer, K. The pyrimidine operon pyrRPB-carA from *Lactococcus lactis. J. Bacteriol.* **183,** 2785–2794 (2001).
61. Michna, R. H., Zhu, B., Mäder, U. & Stülke, J. SubtiWiki 2.0 - an integrated database for the model organism *Bacillus subtilis. Nucleic Acids Res.* **44,** D654-662 (2016).
62. Nicolas, P. *et al.* Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis. Science* **335,** 1103–1106 (2012).
63. Gopal, T., Nagarajan, V. & Elasri, M. O. SATRAT: *Staphylococcus aureus* transcript regulatory network analysis tool. *PeerJ* **3,** (2015).
64. Chang, X. *et al.* EcoBrowser: a web-based tool for visualizing transcriptome data of *Escherichia coli. BMC Res. Notes* **4,** 405 (2011).
65. Ishii, N. *et al.* Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* **316,** 593–597 (2007).
66. Kroger, C. *et al.* An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe* **14,** 683–695 (2013).
67. Kroger, C. *et al.* The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E1277-1286 (2012).
68. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori. Nature* **464,** 250–255 (2010).
69. Toledo-Arana, A. *et al.* The Listeria transcriptional landscape from saprophytism to virulence. *Nature* **459,** 950–956 (2009).
70. Ogunniyi, A. D. *et al.* Identification of genes that contribute to the pathogenesis of invasive pneumococcal disease by in vivo transcriptomic analysis. *Infect. Immun.* **80,** 3268–3278 (2012).
71. Shafeeq, S., Kuipers, O. P. & Kloosterman, T. G. The role of zinc in the interplay between pathogenic streptococci and their hosts. *Mol. Microbiol.* **88,** 1047–1057 (2013).
72. Lysenko, E. S., Lijek, R. S., Brown, S. P. & Weiser, J. N. Within-host competition drives selection for the capsule virulence determinant of *Streptococcus pneumoniae. Curr. Biol.* **20,** 1222–1226 (2010).
73. Siegel, S. J., Roche, A. M. & Weiser, J. N. Influenza promotes pneumococcal growth during coinfection by providing host sialylated substrates as a nutrient source. *Cell Host Microbe* **16,** 55–67 (2014).
74. Wolf, T., Kämmer, P., Brunke, S. & Linde, J. Two's company: studying interspecies relationships with dual RNA-seq. *Curr. Opin. Microbiol.* **42,** 7–12 (2017).
75. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).